A Physiology Clock for Human Aging

Authors: Sergiy Libert[1*], Alex Chekholko[1], Cynthia Kenyon[1*]

1, Calico Life Sciences, South San Francisco, CA, 94080
*cynthia@calicolabs.com
*libert@calicolabs.com

Abstract

Why people age at different rates is a fundamental unsolved problem in biology. We created a model that predicts an individual's age, taking as input physiological traits that change with age in the large UK Biobank dataset, such as blood pressure, blood metabolites, strength, and stimulus-reaction time. The model's Root Mean Square Error of age prediction (RMSE) is less than 5 years. We argue that the difference between calculated "biological" age and actual age (ΔAge) reflects an individual's relative youthfulness and possibly their rate of aging. Validating this interpretation, people predicted to be physiologically young for their age have a lower subsequent mortality rate and a higher parental age at death, even though no mortality data were used to calculate ΔAge. A Genome-Wide Association Study (GWAS) of ΔAge, and analysis of environmental factors associated with ΔAge identified known as well as new factors that may influence human aging, including genes involved in synapse biology and a tendency to play computer games. We identify 12 readily-measured physiological traits that together assess a person's biological age and may be used clinically to evaluate therapeutics designed to slow aging and extend healthy life.

Keywords

Human aging, UKBB, longevity, GWAS, healthspan, biological age, cognition

Introduction

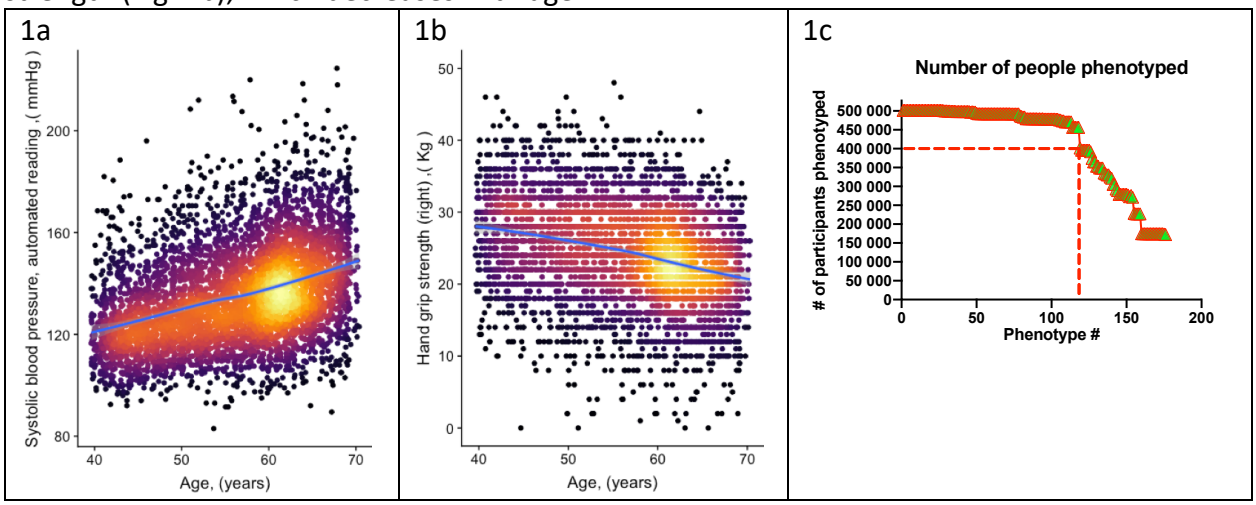
The rate of aging; that is, the rate at which organisms lose physiological fitness and accumulate morbidity, has both genetic and environmental determinants. We all age more slowly than, say, dogs, so genes play a key role, but environmental factors like smoking and exercise influence aging as well.

Previous studies that attempted to identify genes that influence human aging [1–3] have generally considered traits that involve a single parameter, like a lifespan, or parents' age at death, or the presence or absence of age-associated diseases. In this study, we aimed to measure the aging process more directly and systemically. To compile a more inclusive model of human aging, we queried a large, well-annotated human database (the United Kingdom Biobank project or UKBB) comprising phenotypes that together report on multiple organs and physiological systems. We built a mathematical model that considers 121 age-related traits and predicts a biological age for each individual. By comparing predicted biological age to the real age of the person, we can identify individuals that might be aging slower or faster than average, and then analyze their genes and environments to look for candidate determinants that influence aging. In addition, we describe twelve physiological traits that together could measure biological age in longitudinal clinical trials for interventions that slow human aging and increase healthspan.

Results

Phenotypes contributing to the model for age prediction

The UK BioBank is an unparalleled resource comprising health, environmental, and genetic information for over 500,000 people. Traits that change with age underlie models of age prediction. To identify age-dependent traits that reflect a decline in physiology; that is, traits we "care about" in everyday life, we used linear regression for every UKBB parameter relative to the age of the participants (Sup. Table 1 a, b) and identified those that had a non-zero slope with statistical significance better than 10^{-3} (Sup. Table 2 a,b). Examples of these age-related phenotypes include systolic blood pressure (Fig. 1a), which increases with age, and hand-grip strength (Fig. 1b), which decreases with age.



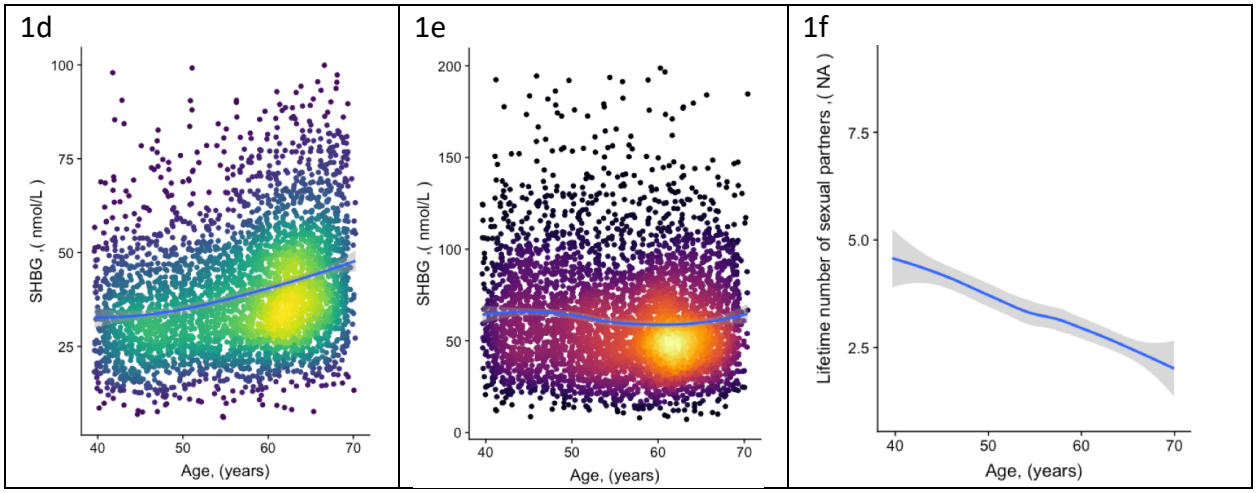


Figure 1. a) Systolic blood pressure (UKBB field ID# 4080) and b) hand-grip strength (UKBB field ID# 47) of a random set of 10,000 female UKBB participants is plotted against their age (UKBB fields ID# 4080, 47(should 47 be there?)); c) Number of age-sensitive phenotypes plotted against the declining number of people in whom these phenotypes were measured; d) sex hormone binding globulin concentrations of a random set of 10,000 males (UKBB field# 30830) (d) and females (UKBB field# 30830) (e) UKBB participants are plotted against their age (UKBB field ID# 30830); f) Average number of lifetime sexual partners is plotted against age of UKBB participants (UKBB field ID# 2149), grey area denotes 99% confidence interval. Color of dots on the plot represents relative density of dots in the area.

Although the UKBB contains ~500K individuals, not all age-dependent traits were measured in each person. For example, MRI brain scans were only measured in ~50,000 people, so MRI data cannot be used to estimate the age of the remaining participants. To optimize the number of traits we could incorporate as well as the number of people whose data could be included, we plotted the number of traits examined as a function of cohort size. From this curve (Fig. 1c), we estimated that ~120 +/- 15 phenotypes would comprise an optimal inclusion threshold, as the number of people evaluated declined dramatically for larger numbers of phenotypes.

We chose to analyze male and female aging separately, as many age-dependent physiological phenotypes differ between genders. For example, the parameters “age at which first facial hair appeared”, “age at menopause” and “degree of pattern balding” are gender specific. Moreover, shared phenotypes can actually have an opposite dynamic in males versus females. For example, an increasing plasma concentration of sex-hormone binding globulin (SHGB) is one of the best predictors of age in males (Fig. 1d); however, its plasma concentration in females stays nearly constant, or even has a tendency to decrease (Fig. 1e).

In building a model for age prediction, it is important to exclude traits whose age-dependence likely reflects changing cultures or sociological norms across generations rather than intrinsic physiological decline. For example, a good predictor of age (p-value < 10^{-52}) is a lifetime number of sexual partners (Fig. 1f). While sexual activity and fertility have been linked to human aging and longevity [4], here, the correlation is likely influenced by evolving social norms in Britain. Similar traits included “how many siblings do you have”, or “how long have you lived in your current house”. Additionally, we also found some seemingly pure biological measurements to be derived using age as a parameter. For example, BMR (Basal Metabolic

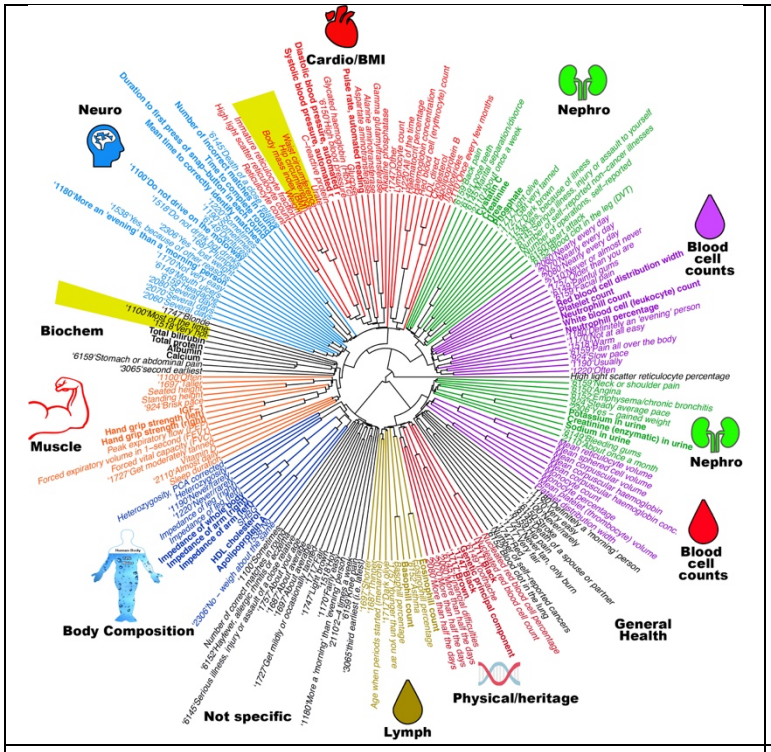
Rate) is an excellent predictor of age (p-value < 10^{-255}). However, BMR was not measured directly, but rather was calculated using a formula containing height, weight, gender, and age itself. For this reason, we investigated and carefully considered each age-dependent trait independently, attempting to satisfy 3 broad criteria: a) the trait should not reflect social norms and structures; b) the trait should not be a function of passing time (how long have you been drinking green tea?); and c) the trait's value should not be calculated using a person's actual age. We strove to utilize pure biological and physiological parameters only. The full list of the age-related traits we selected, and our justifications for doing so, are detailed in supplementary tables S2 c, d.

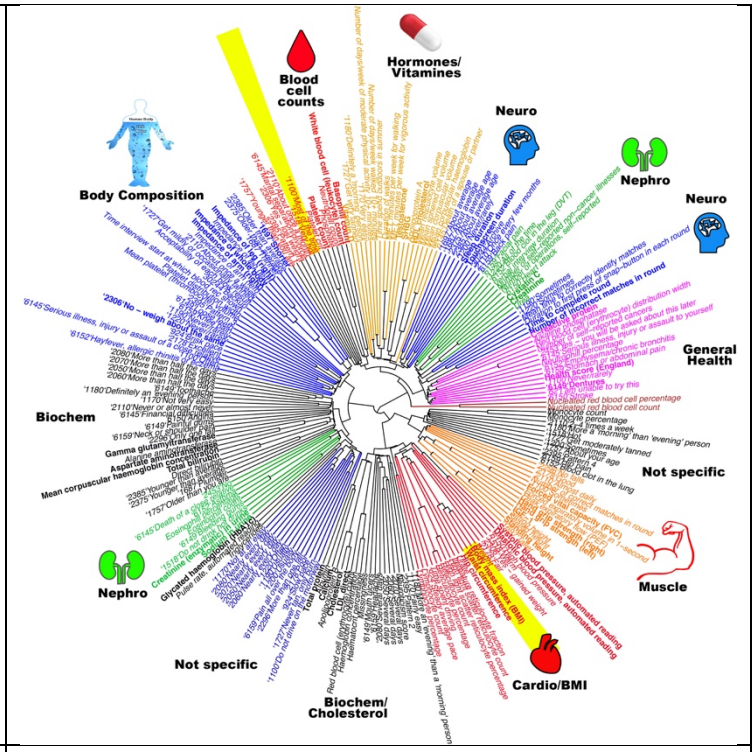
A mathematical model to predict age

We identified a list of biological phenotypes that change monotonically with age across the population. We confirmed that these selected phenotypes likely have the same trajectory within individuals if one is to measure them longitudinally. As expected, many of these age-dependent phenotypes are correlated, such as left- and right-hand grip strength. These correlations necessitate additional data normalization, so that certain biological features are not represented multiple times. To investigate the degree and pattern of correlations among the age-dependent traits within each individual (Supp. Table S1), each phenotype was normalized (see supp. methods) and Euclidean distances and Pearson correlations were calculated for each pair of phenotypes. Correlations between phenotypes are displayed as dendrograms (fig. 2). In a dendrogram, highly correlated phenotypes cluster together, such as "BMI"-"Weight"-"Waist Circumference", or "Cholesterol"- "LDL". Interestingly, this analysis also identified intriguing correlations that await mechanistic explanation, such as "I drive faster than the speed limit most of the time" (id# 1100) with "I like my drinks very hot" (id# 1518) (fig. 2a,b; marked with yellow shadows).

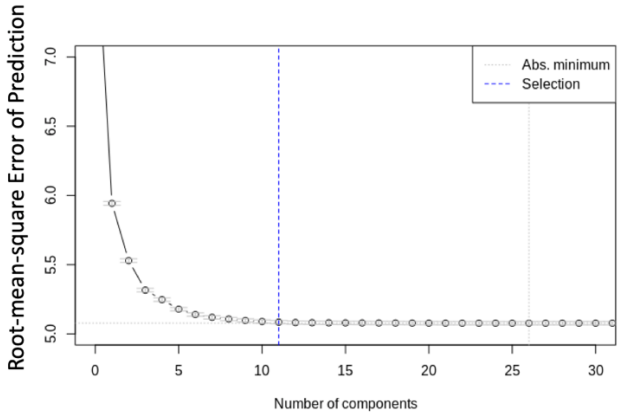
Notably, most of the clusters appeared to be enriched for phenotypes associated with a specific organ or physiological system. For example, the cluster that contains "Creatinine", "Urea", "Cystatin-C" and "Phosphate" likely reflects kidney function; whereas the cluster that contains "Systolic blood pressure" and "Diastolic blood pressure" likely reflects cardiovascular function (fig. 2). Again, in some cases, it is not intuitively obvious why certain physiological traits do or do not cluster with one another. Thus, this dendrogram might be a valuable data source for future hypothesis generation and exploration.

2a (female)	2b (male)

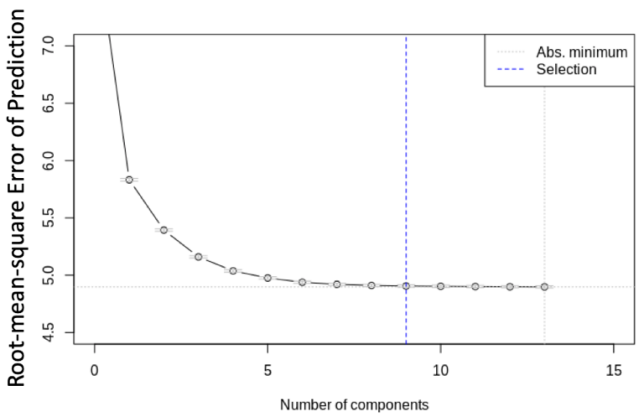




2c (female)



2d (male)



2e (female)

Parameter	Weight
Systolic blood pressure, automated reading	13.21
Forced expiratory volume in 1-second (FEV1)	9.25
IGF-1	8.82
Urea	8.09
Cystatin C	8.05
Mean time to correctly identify matches	6.63
Diastolic blood pressure, automated reading	5.99
Sodium in urine	5.45
Glycated haemoglobin (HbA1c)	5.28
Dentures	5.20
Alkaline phosphatase	5.14
Hand grip strength (left)	4.31
Calcium	4.17

2f (male)

Parameter	Weight
Systolic blood pressure, automated reading	12.22
SHBG	11.79
Forced expiratory volume in 1-second (FEV1)	11.26
Waist circumference	9.10
Cystatin C	7.84
Testosterone	7.33
Mean time to correctly identify matches	7.21
Dentures	6.77
Touchscreen duration	6.08
IGF-1	5.70
Body mass index (BMI)	5.64
Sodium in urine	5.55
Urea	5.43

2g (female)

2h (male)

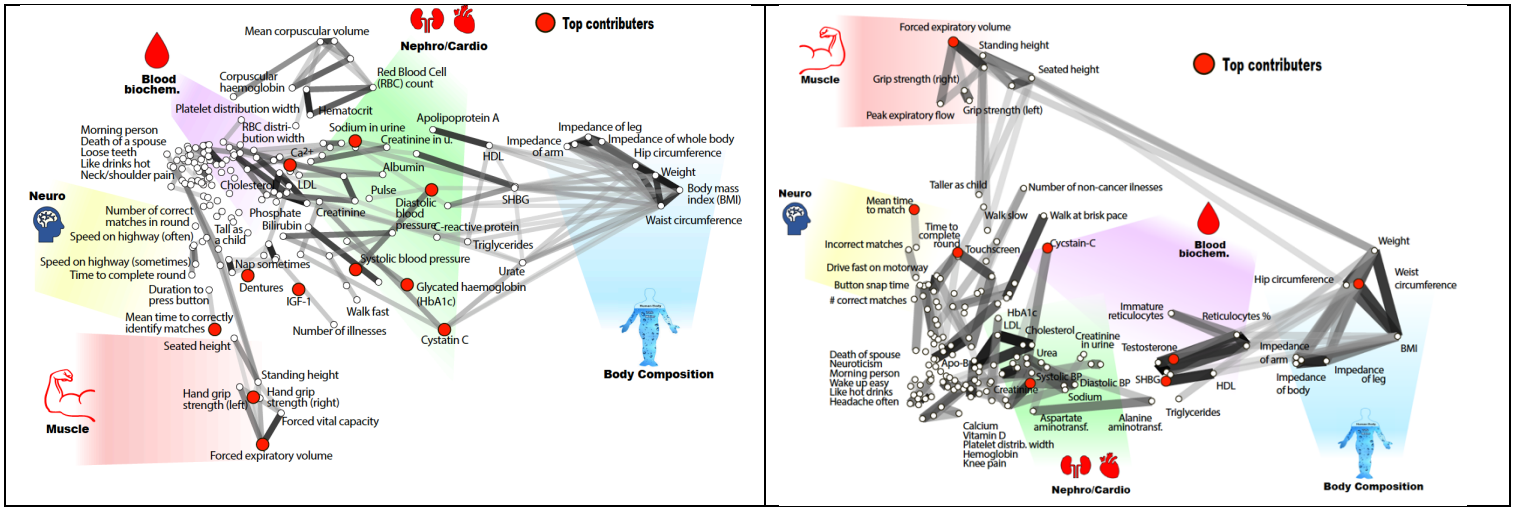


Figure 2. Age-dependent phenotypic clustering. Dendrogram plots of age-dependent female (a) and male (b) phenotypes selected for age prediction. Numbers in the name of "rays" represent UKBB ID numbers for multichoice questions (see supplementary table 1 or the UKBB website), followed by the answer. Major clusters were colored and subjectively assigned a name that reflects a possible biological function of the cluster; The number of principle components included in the PLS (Projection to Latent Structures) model to predict age vs root mean square error of the predictions is plotted for females (c) and (d) males; e) the top phenotypes with the highest "weights" in the age-predicting PLS model are listed for (e) females and (f) males. Phenotypes shaded in green are shared between sexes, red are specific to females, and blue are specific to males. It should be noted that all phenotypes were used for both sexes, and this shading reflects only the position in list of top-13 traits; g) list of phenotypes used to predict age of females and (h) males projected on 2D space using Pearson correlation as the distance measure. The degree of correlation is also depicted by grey lines; the darker the shade, the stronger the correlation. It must be noted that the distortion in positioning is an inevitable consequence of projecting high-dimensional data into 2D space. As before, groups of related phenotypes were subjectively assigned a name that likely depicts their physiology, and phenotypes with the highest weight in the PLS model were depicted by red dots.

To construct a model that predicts age, we chose to employ the "Partial Least Squares regression, or Projection to Latent Structures" (PLS) method, as this method would allow us to identify components that capture the most age variation. PLS modelling, however, is non-tolerant of missing values. In the UKBB dataset that we selected, over 60,000 participants (~15%) lacked at least one measurement for one of the phenotypes. To mitigate against excessive imputation, any individual missing more than 15 datapoints was excluded from the study. In females, the number of selected participants decreased from 222,111 to 215,949 (~2.7% loss), and in males from 188,609 to 183,715 (~2.6% loss). The rest of participants with missing data had the values imputed and then scaled using the KNN algorithm of the impute R package (described in Methods).

We wondered how many PLS components (each derived from UKBB phenotypes) were required for optimal prediction of chronological age. To do so, we built a series of age-prediction models using increasing numbers of these components: the first model using just component #1; the second using components #1 and #2, and so on. At each step, we measured the root-mean-square error of the age prediction and assessed its decline using the R function

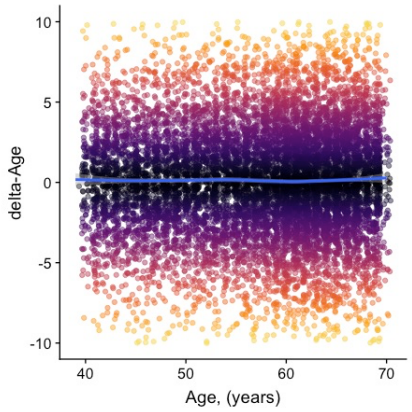
"selectNcomp" (fig. 2c, d). This analysis indicated that describing female aging dynamics requires 11 independent components and that describing male dynamics requires 9 independent components. Adding additional components did not result in further improvement of the corresponding models. Thus, the models to predict age using UKBB phenotypes were built using R function "plsR", with 9 and 11 components for males and females, respectively, using the Cross-Validation function (CV) to prevent overfitting.

It was interesting to ask which individual age-sensitive phenotypes were most valuable for age prediction. Because many phenotypes contribute to more than one PLS component, to do this, we deconstructed each of the PLS components and summed up the absolute values for phenotype coefficients across all the components. In this way we derived a "weight" metric for each phenotype used to predict age. The top twelve phenotypes with the highest weights; that is the phenotypes that together can best predict biological age, are shown on figures 2e and 2f. Most of the top phenotypes were shared between males and females and reflected specific physiological systems; for example, systolic blood pressure (which likely correlates with cardiovascular health), forced expiratory volume (a likely correlate of pulmonary and cartilage/bone health), urea and cystatin C levels (likely correlates of kidney health) and mean time to identify matches correctly (a likely correlate of cognitive health).

Inferred (biological) age predicts all-cause mortality better than passport age

Using the physiological phenotypes listed in tables S2a, b and the PLS modelling described above, we were able to predict individual age with a root mean square error of 5 years (figures 2e, f). There are many sources for a discrepancy between predicted biological age and chronological age. These include statistical noise, unknown variations in life histories of UKBB participants, limited accuracy of certain measurements, and insufficient numbers of relevant measurements. However, it is also possible that part of the discrepancy between chronological age and predicted age is due to some individuals' aging slower and others aging faster than average. While it is not possible to calculate any one individual's rate of aging with certainty without longitudinal measurements, at the population level one can generate plausible hypotheses from cross-sectional analysis, hypotheses that can be tested rigorously by longitudinal studies incorporating deep phenotyping in the future. Thus, to infer the biological age from this cross-sectional data, we calculated a value we call ΔAge for each individual. ΔAge is the person's age as predicted by the model minus his or her chronological age, normalized such that average ΔAge for the whole population at each age is zero. Similar approaches have been used before; for example, using DNA methylation patterns to derive predicted age [5], or using facial images and computer vision to predict age and identify potentially "fast agers" and "slow agers" [6]. ΔAge is negative if an individual is predicted to be younger than they are and positive if an individual is predicted to be older than they are. Note that the ΔAge parameter carries no information about the person's actual chronological age—it simply represents the difference between his or her actual and biological age (fig. 3a).

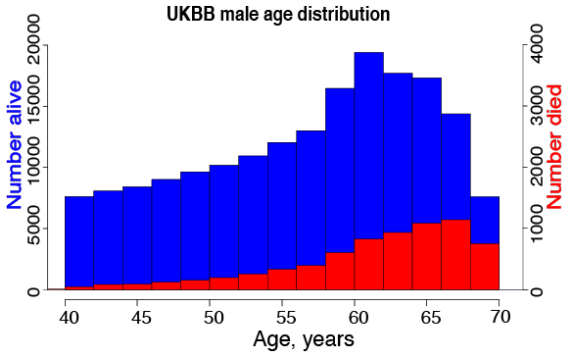
3a



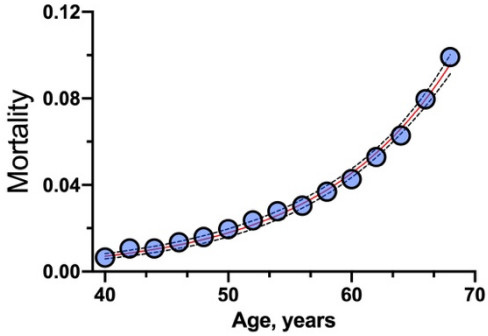
3b

red-died

blue-alive

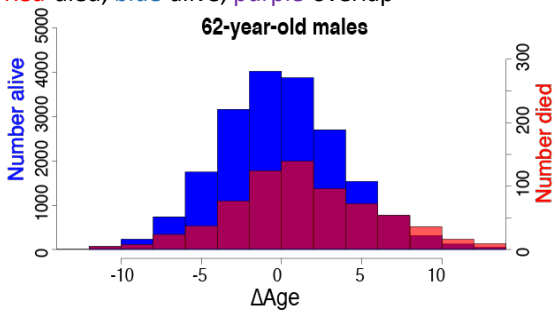


3c

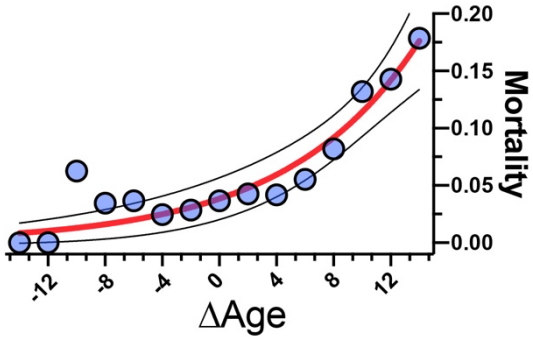


3d

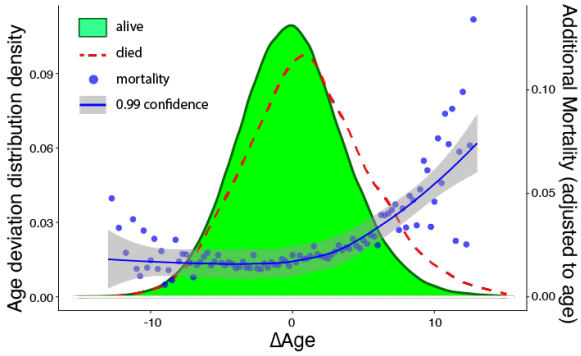
Red-died; blue-alive; purple-overlap



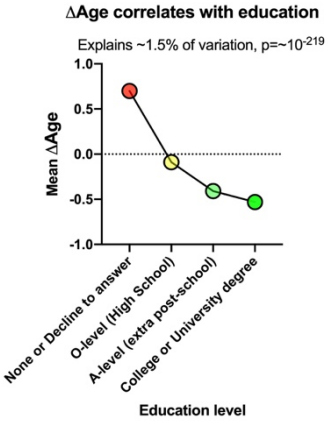
3e



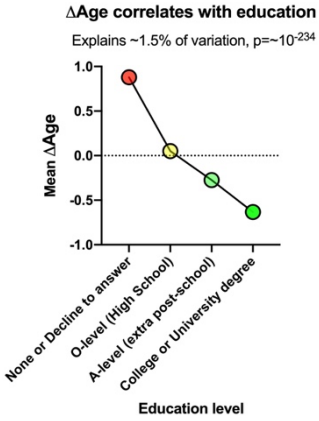
3f



3g (male)



3h (female)



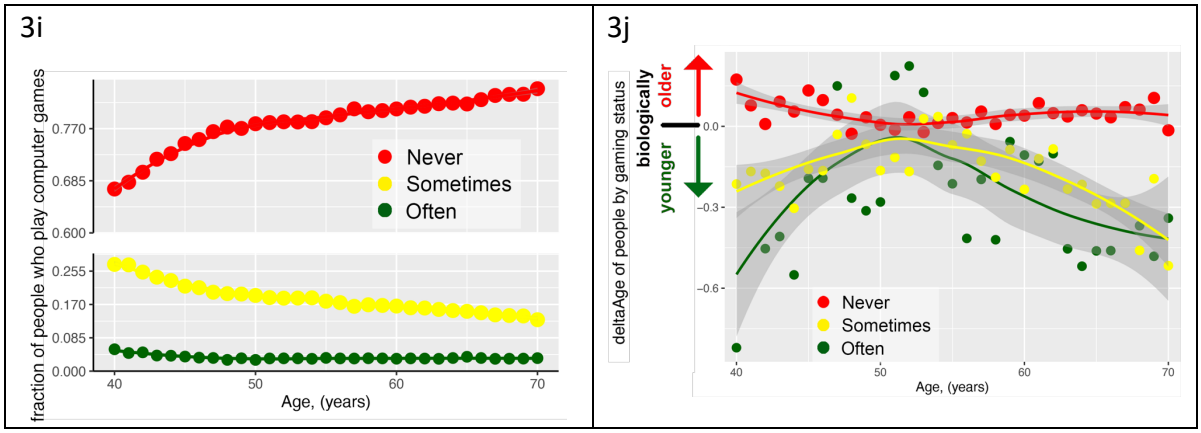


Figure 3. ΔAge has biological meaning. a) delta-age (ΔAge, predicted biological age minus chronological age) is plotted against chronological age for a random subset of 10,000 UKBB participants. Note that there is no correlation between age and ΔAge; b) histogram of age distribution (blue), and death distribution (red, right y-axis) is presented for UKBB males; c) mortality of UKBB male participants vs their age is platted, note the classical exponential (Gompertzian) shape. Blue dots are actual data, the red line is an exponential fit, and the black dash line is 95% confidence interval; d) histogram of the ΔAge distribution (blue), and death distribution (red, right y-axis) is presented for UKBB males of 62 years of age only; e) mortality of 62-year-old males is plotted against their ΔAge. Blue dots are actual data, the red line is an exponential fit, and the black dash line is 95% confidence interval. Once again, note the classical exponential (Gompertzian) shape with ΔAge, even though all the subjects are the same age; f) distribution of ΔAge for all the people in UKBB (all ages and all genders, green shape), the distribution of ΔAge for people who died within 5 years after enrolling in the UKBB (red line) is shown for comparison; note a shift of the deceased distribution to the right, towards larger ΔAge (predicted older on average). The mortality penalty due to ΔAge is plotted as blue dots (left y-axis), the exponential fit of these data is presented as a blue line, and the 99% confidence interval as a grey shade; The average ΔAge is plotted for UKBB males (g) and females (h) against their highest education (qualification) achieved; i) the fraction of people who play computer games "sometimes" (yellow dots), never (red dots), and "true gamers", who play computer games "often" (green dots); j) average Δage of people at different ages separated by their computer gaming habits (see 3i). Note, people who play computer games "often" are biologically younger than people who play computer games "sometimes", or "never".

In the classical paradigm of aging described by Gompertz, mortality rates increase exponentially with time, with the human rate doubling every 8 years [7]. In the UKBB dataset that we analyzed, 8883 males (fig. 3b) and 5668 females died within 5 years of their initial test-center attendance. The distribution of these deaths among UKBB participants has a classical "Gompertzian" shape, with mortality rates increasing exponentially with a 7.7-year doubling rate for both males and females, as can be seen in figure 3c. In this simplistic model, where mortality depends on age only, everyone of the same age has exactly the same chance of dying. However, we found that using the ΔAge parameter (predicted biological age minus passport age), we could predict death among people of the same age more accurately. For example, consider 62-year-old males. In this cohort (fig. 3d), individuals with negative ΔAge values were predicted to be biologically younger than 62, and vice versa. Among these males, several hundred died within 5 years of their enrollment. Plotting the average mortality for every ΔAge bin in this stratification of 62-year-olds yielded a Gompertz-like mortality distribution.

Remarkably, one year of ΔAge produced the same effect on mortality rate as did one year of chronological age: the rate of mortality doubling was 7.2 years of ΔAge (biological age), which is very similar to the 7.7-year rate of mortality doubling for chronological years among UKBB males. It must be reiterated that death data were not considered at any point in building the model of biological age or ΔAge derivation, and that ΔAge has no correlation with real age (fig. 3a). Yet the ability of ΔAge to predict death with similar accuracy as real age is robust across genders and ages. In fact, it can even be observed when people of all ages are combined together (fig. 3f). Although ΔAge itself is age-independent, it follows that the combination of chronological age plus ΔAge produces a more accurate prediction of mortality risk than does chronological age alone, a fact of potential interest to actuaries.

The fact that ΔAge correlates with mortality rates suggests that it would also correlate with medical history, as most deaths are caused by disease. Indeed, this was the case; for example, ΔAge correlated with "number of treatments/medications taken" (UKBB field #137). Interestingly, we also saw a strong correlation of ΔAge with father's age at death (p-value = $1.9*10^{-43}$ for females, and p-value = $3.9*10^{-31}$ for males), and mother's age at death (p-value = $1.1*10^{-68}$ for females, and p-value = $1.3*10^{-32}$ for males), with subjects predicted to be biologically younger having parents die at an older age. Lifespans of parents and offspring have previously been reported to be correlated [8]. Again, this finding reinforces the notion that ΔAge captures information about the aging process beyond just age.

A central goal of current aging research is to find interventions that improve the quality of human aging and extend the period of vigor and health. The age-dependent traits measured in the UKBB include those we care most about, such as strength and cognition, the traits that together define our quality of life; therefore, measuring a ΔAge parameter created from these traits might best capture biological aging. In a real-world clinical trial for healthspan, measuring over 120 physiological parameters would be expensive and unfeasible; thus, we investigated if a smaller subset might be an adequate surrogate for a full ΔAge model. We chose to build a simple linear model using the top 12 parameters (parameters with the highest weight) identified by our PLS model (fig. 2e,f). Among these top 12 parameters, 7 were the same for males and females; namely, systolic blood pressure, forced expiratory volume in 1 second, cystatin C concentration, mean time to correctly identify matches, presence of dentures, IGF-1 levels, and sodium in urine. In addition to these 7 core measurements, additional 5 were gender specific: calcium, left handgrip strength, alkaline phosphatase, glycated hemoglobin, and diastolic blood pressure for women, and sex-hormone-binding globulin, waist circumference, testosterone, duration of touchscreen, and BMI for males. Using just these 12 parameters it is possible to build a linear model that captures 86.9% of information carried by a full model ΔAge. In fact, ΔAge calculated using these truncated 12-parameter models were able to predict additional to age mortality in males and females with p-values $< 10^{200}$.

Environmental factors that influence biological age

Our mathematical model, constructed from physiological traits, allowed us to identify environmental factors that associate ΔAge. Human lifespan is known to be positively associated

with personal wealth [9] and is negatively associated with smoking or excessive drinking. To investigate whether ΔAge has similar associations, and possibly to identify new environmental parameters that influence aging, we calculated the correlation of ΔAge with every parameter available from UKBB, asking, for example "Does smoking correlate with increased biological age"? (supplemental tables S3a, b). Correlations with p-values lower than 10^{-5} were considered statistically significant. (This cutoff was calculated to correct for multiple testing with the base error of 0.05 and 3300 independent tests.) It was reassuring to see that all the parameters that had been used in ΔAge derivation (such as blood pressure or hand-grip strength) had strong and highly significant (p-values less than 10^{-200}) correlations with ΔAge. Interestingly, we saw a strong association of ΔAge with age-dependent biological phenotypes that were not used in the model to predict ΔAge due to the low number of people who underwent the assessment. For example, heel bone density (UKBB field # 3148) and Thalamus volume (UKBB field # 25011) both had strong associations with ΔAge (p-values were $6*10^{-11}$ and $6*10^{-10}$). These and other phenotypes with strong ΔAge correlations might be useful parameters to consider when building biological clocks in the future.

The list of purely environmental factors that correlate with ΔAge is presented in supplemental tables S3c, d. As predicted, we found that wealth correlated with a more youthful ΔAge. For example, parameters such as "home location" (UKBB filed id# 20075), "place of birth" (UKBB field id# 129), "Townsend deprivation index" (UKBB field id#189), and "total income" each had a strong, significant correlation with ΔAge (Tables S3). Likewise, as expected, "smoking" and "exposure to smoke" (field ids# 20161, and 20162), and certain spirits (Supp. Item #1) had a positive correlation with ΔAge.

Interestingly, the single most significant (p-value<10^{-200}) non-biological parameter to correlate with ΔAge in both males and females was "Qualifications" or the level of education achieved (UKBB field id# 6138). Every additional level of education achieved was progressively associated with increased "youthfulness" (fig. 3g,h), and the size effect of the education (-1.51) was much greater than that of wealth (-0.81) or place of birth (-0.13).

Certain leisure and social activities were also correlated with ΔAge. The amount of TV watching (UKBB filed# 1070) was positively correlated with ΔAge in both males and females, whereas time spent outdoors (UKBB filed# 1050) for males, and DIY projects (UKBB filed# 2624) for females were correlated with younger ΔAge. Intriguingly, the second strongest behavioral traits that associate with ΔAge (which were not part of the model) was the "frequency with which people play computer games". This is a novel association, and one that is less likely to reflect socioeconomic status, as access to computer gaming is cheap and widely available. Playing computer games associates with youthfulness (fig. 3i,j, Supp. Item #2), with a size effect of -2.2 and p-value of 4*10^-8. This association is equally strong if "age" is factored out from the regression, indicating that generational changes in leisure activities do not explain this association.

Overall, the correlation analysis of ΔAge with environmental variables confirmed already known associations, such as the positive associations of wealth, sports, and social engagements and the negative association with smoking. An interesting finding was the highly significant association and large effect size of ΔAge with computer gaming and formal education, which,

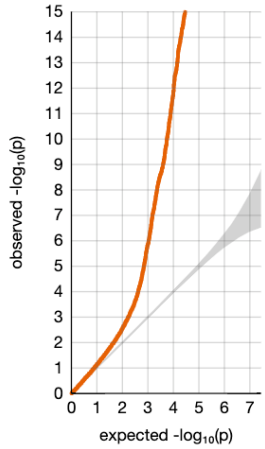
however, still underscores the important role that socio-economic factors play in human health and successful aging [9,10]. To our knowledge this is the first reported association of computer game playing habits and biological age in humans. Understanding this association might help to better understand age-dependent changes in the nervous system and strategies to counteract them.

Genes that might influence biological age

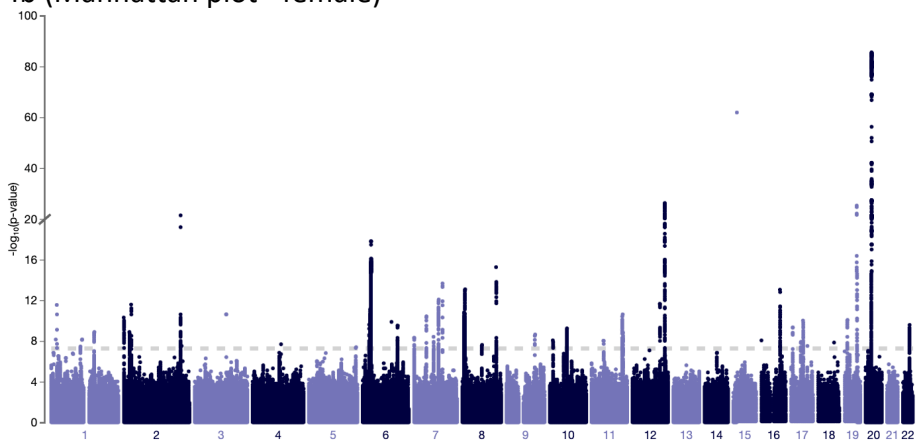
To identify genes that might influence human aging, we performed a genome-wide association study (GWAS) using linear models separately on males and females. Analysis was performed using Hail-0.2.38 software running on a Debian Linux cluster with 72 cores. To account for the influence of environmental factors on ΔAge, we corrected for smoking history, education, income level, Townsend deprivation index (multiparametric measure of socio-economic status), levels of air pollution, and geographic location of the primary house by including these variables as covariates. GWAS was performed on men and women of white British decent only, and to decrease the noise due to the possible population substructure, we included first forty (40) principal components derived from genetic data of the participants and genotyping batch as covariates during the analysis. Additionally, recent work [14] identified irregularities in variant frequencies of certain SNPs in the UKBB dataset, and these questionable SNPS were excluded (Supp. table S5).

The GWAS results were analyzed and genome-wide significance of associations was determined using the PheWeb engine (version 1.1.14). The nearest genes for all identified loci were called using the GRCh37 human reference genome. QQ- and Manhattan plots for male and female models are presented on figures 4a-d, and summary statistics for the GWAS analyses are available as supplementary tables S4. Briefly, for male ΔAge GWAS, λ_{gc}=1.2005, and LD regression intercept= 1.0213+/-0.0083. For female ΔAge GWAS, λ_{gc} =1.2531, and LD regression intercept= 1.0285+/-0.0119.

4a (QQ plot-female)



4b (Manhattan plot - female)



4c (QQ plot-male)

4d (Manhattan plot - male)

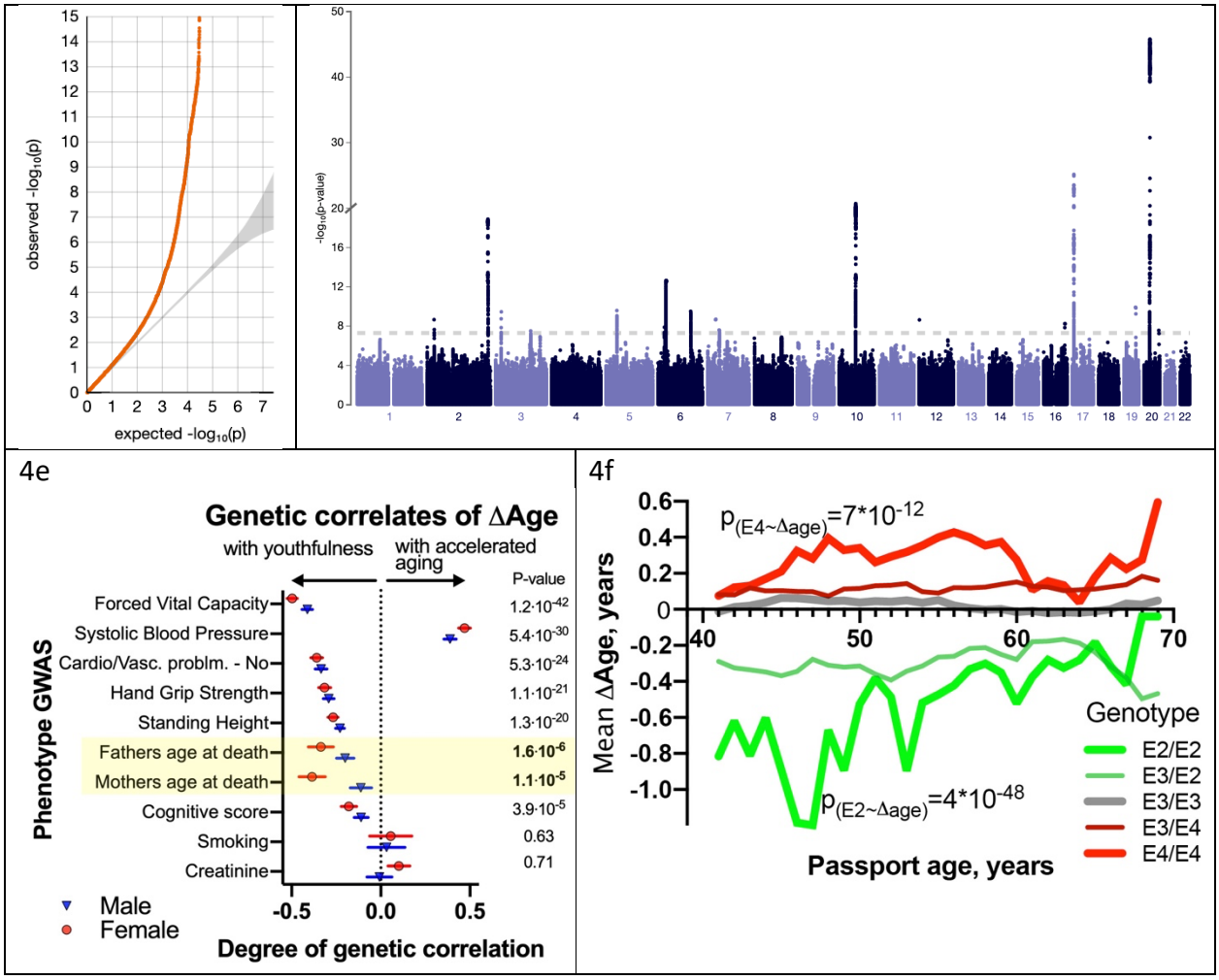


Figure 4. Genetic analysis of ΔAge. a, c) Quantile-Quantile plots for female and male -log10 p-values; b, d) Manhattan plots from genome wide association analysis of female and male ΔAge; e) Genetic correlation of ΔAge determination with other GWAS performed and reported by UKBB consortium. Note the strong genetic relation between ΔAge and parental age at death; f) effect of *APOE* alleles on average ΔAge plotted across different ages. Beneficial *APOE2* alleles are in green, and detrimental *APOE4* alleles are in red.

These ΔAge genome-wide association studies identified 84 loci in females (fig. 4a, b, table. S4) and 47 loci in males (fig. 4c, d, table. S4) following correction for multiple testing (p-value < 10^{-6}), thus potentially influencing whether a person was biologically older or younger than average for their age. The list of these loci is available in supplementary tables S4, and the list of genes closest to those loci is presented on figure 5b and supplementary table S4. Six loci, whose closest genes are *CST3, HLA-DQ, APOE, GCKR, PIK3CG*, and *SMARCA4,* were identified independently in both males and females. This overlap is significant, $p < 10^{-8}$. Interestingly, a total of 15 female hits had p-values of < 10^{-5} in males, and a total of 9 male hits had p-values of < 10^{-5} in females; thus, further analysis might indicate a role for some (but likely not all) of these genes in both sexes, see supplementary figure 5. Consistent with the concept of underpowering, when we combined male and female ΔAge data, the GWAS analysis identified

12 more loci in addition to those already identified in sex-specific studies (table S4, and figure 5b).

Many of the genes identified in our analysis had been associated previously with classical diseases of aging. For example, the *HNF1A* (hepatocyte nuclear factor 1 homeobox A) locus (top SNP – rs1169284, ΔAge association p-value = $3.0*10^{-23}$) had been associated with diabetes [15] and cancer [16]. The *APOE* (apolipoprotein E) locus (top SNP – rs7412, ΔAge association p-value = $4.4*10^{-33}$) had been associated with Alzheimer's disease and coronary heart disease [17]. It is possible that people who carry risk alleles for age-related disease have a less-robust physiology (higher ΔAge) due to the disease itself, even though their overall aging might have been otherwise unaffected. To investigate this possibility, we repeated our analysis while excluding all the people who had been diagnosed with age-related diseases; specifically, people diagnosed with diabetes, cancer, asthma, emphysema, bronchitis, chronic obstructive pulmonary disease (COPD), cystic fibrosis, sarcoidosis, pulmonary fibrosis tuberculosis, any vascular or heart problems (such as high blood pressure, stroke, angina or heart attack) or anybody with a history of allergic rhinitis. These strict exclusion-criteria decreased the number of the people in the study by almost 50%. Interestingly, genome-wide association analysis of this smaller but healthier cohort almost completely confirmed genetic associations for ΔAge established using the larger, less healthy, cohort (supp. table 4). The top 20 hits were all confirmed, albeit with less significant p-values. The full GWAS summary is presented in supplementary table 4. These findings suggest that most of the genetic signal associated with ΔAge is coming not from a few susceptibility alleles for specific diseases but rather from alleles that describe and possibly drive fundamental processes that change with age; that is, possibly with aging itself. Conversely, this analysis also allowed us to identify genes that are narrowly responsible for certain diseases that manifest similarly to accelerated aging. For example, the GCKR (glucokinase regulatory protein) locus has a strong association with ΔAge (p-value=$8*10^{-12}$); however, the association disappeared when we excluded people who had been diagnosed with diabetes, demonstrating that mutations in GCKR cause a specific disease and do not have a detectable effect on the healthy cohort. However, some caution should be exercised while interpreting analysis of this smaller "healthier" subpopulation. It is possible that due to the decreased sample size, certain hits disappeared not due to diseases but due to decreased statistical power, which could result in false negatives. Conversely, it is possible that some people had undiagnosed or subclinical disease, and therefore that some of the surviving hits are due to disease, thus resulting in false positives. Additionally, some of the associations might be false positives due to Collider bias. With these caveats in mind, however, it is possible that among the GWAS hits that disappear when diseased individuals are removed from the dataset are new disease-susceptibility genes, while those that persisted are more likely responsible for general aging.

To characterize our GWAS hits further, we performed Linkage Disequilibrium (LD) score regression analysis [18]. This analysis allows us to estimate the SNP heritability of complex traits and calculate the genetic correlation between different phenotypes. Overall, this analysis included 1,293,150 unique SNPs that had an allele frequency higher than 0.01. The calculated genomic inflation factor (lambda GC) for these studies was 1.25. We found that total genetic

heritability (h^2) of ΔAge was ~11% (0.108+/-0.009) for females and ~10% (0.096+/-0.008) for males, which is similar to the genetic heritability estimated previously for human longevity [2,19].

We found that the genetic signature (Manhattan plots) of some of the components used to calculate ΔAge was correlated to the genetic signature of ΔAge itself (Fig. 4e). For example, GWAS of Forced Vital Capacity (FVC) has a correlation with ΔAge GWAS of 0.49 +/- 0.02 (p-value=$5*10^{-65}$). Such a genetic correlation is expected and serves as an independent verification that FVC and FEV play major roles in ΔAge determination, as we established earlier (Fig. 2e, f). This correlation suggests the hypothesis that diminishing FVC/FEV might be a driver of aging. For example, reduced FVC/FEB might influence the aging of multiple tissues by restricting their supply of oxygen, or alternatively, age-dependent biochemical processes that support the lung might support other tissues as well. Contrary to FVC/FEV, other physiological parameters, such as blood creatinine levels, which are explicitly used in ΔAge derivation, had no genetic correlation with ΔAge (0.1+/-0.07, p-value=0.1). This suggests that such traits are likely biomarkers and not drivers of aging. Certain metabolic parameters had been correlated to mortality before [20], but it has been an open question if those metabolites directly drive aging and mortality.

Interestingly, the genetic signature (Manhattan plot) of ΔAge has a strong similarity to genetic signature obtained through GWAS for "Father's age at death", and "Mother's age at death". The genetic correlation of GWAS for parent's age mortality with GWAS for offspring's ΔAge was 0.39+/-0.03, p-value=$1*10^{-7}$ for females and 0.2+/-0.05, p-value=$3*10^{-5}$ for males, even though mortalities of subjects or parents were not part of the model. These genetic correlations once again demonstrate that ΔAge carries information about aging and longevity, even though its values are derived from cross-sectional physiological data and independently of lifespan.

To investigate if there are distinct pathways that have an influence on human biological age (and possibly the rate of aging), we performed GeneOntology analysis of GWAS hits (combined male and female genetic loci). Five enriched pathways were identified in this analysis. The top enriched category (GO:98815, *modulation of excitatory postsynaptic potential*) was enriched ~18 fold over the expected by-chance reference, with multiple-testing-adjusted p-value of 0.046. This category stood out by the enrichment strength (~18 fold), as the second-best enrichment category was enriched only ~3 fold (*response to oxygen-containing compounds*). The full list and statistics are presented in the Supp Item #3.

The gene with the most significant association with ΔAge was CST3. CST3 encodes cystatin-C, a metabolite whose concentration increases monotonically with age. Levels of cystatin-C are routinely used to evaluate kidney health; moreover, elevated levels of this metabolite had been linked to elevated risk of cardiovascular disease (CVD) [21], risks of cancer [22] and neurodegeneration [23]. However, in a Mendelian Randomization Study [21] it was shown that while levels of cystatin-C predict CVD well, SNPs that robustly alter expression of cystatin-C do not associate with CVD. Similarly, we found that if we calculated ΔAge using all the inputs as before except excluding cystatin-C levels, the subsequent GWAS analysis of that model yielded nearly identical results, except that the locus of *CST3* (the gene encoding cystatin-C) was no

longer a hit. Combined, these data suggest that cystatin-C is a “marker” and should not be considered a therapeutic target.

The association of *CST3* locus with ΔAge was completely eliminated when a single physiological trait (level of cystatin 3, the product of *CST3*) was excluded from the model. We expect that a gene that fundamentally influences the overall rate of aging making our rates of aging different from that of dogs) ideally would affect all age-related traits and would not depend on a single phenotype. Therefore, we sought to identify all the GWAS hits that were eliminated when we left out a small set of highly correlated phenotypic traits. We built 10 “drop-out” models, in which physiological traits comprising one phenotypic cluster were excluded from the model. These phenotypic clusters (shown in figures 2a,b,g,h and detailed in sup table 2) were generated mathematically by hierarchical clustering. By eye, these clusters appeared to represent muscle function (drop-out model 1), body composition (2), kidney health (3), cardio health (4), blood cell composition (5), blood biochemistry (6), neuro-psychiatric phenotypes (7), lipid metabolism (8), physical attributes (9) or general health (10). As expected for a true model of aging, ΔAge values remained robust among all the drop-out models, as demonstrated in Supp. Item #4.

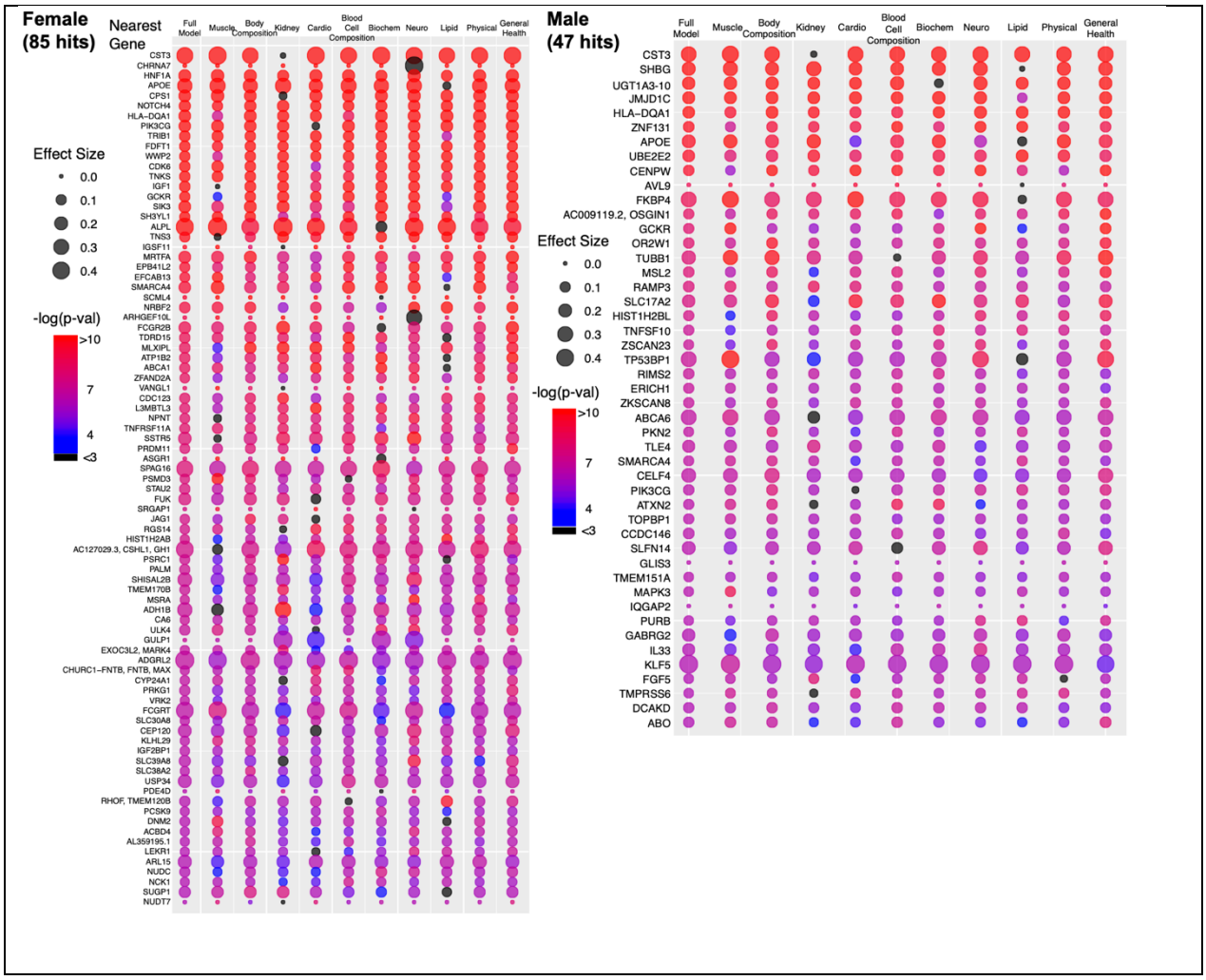


Figure 5. Cluster drop-out models. The list of nearest genes to identified GWAS loci that associate with female and male ΔAge in the full model. Each hit is presented as a bubble, colored according to the significance of association of the locus with ΔAge, with size representing the effect size of the top SNP in the locus. The full summary is presented in Supplementary table 4.

To evaluate each candidate GWAS hit, we took the best SNP from each locus from the full male or female analysis (above) and tested its association with ΔAge computed using each of the 10 drop-out models. The bubble-plots in figure 5b represent the effect size of each of these SNPs (via bubble size), and the associated p-value (via color). As predicted, the GWAS hit CST3 lost all statistical significance when traits within the cluster containing cystatin C (dropout model 3, kidney health cluster) were excluded from the model. In contrast, some GWAS hits never dropped out, and these remained candidates for fundamental determinants of ΔAge and potentially the human rate of aging.

Discussion

In this study, we constructed a model predicting human age based on a wide variety of age-dependent parameters that together report on the major organs and physiological systems of the body. When this model overpredicts or underpredicts one's age, the discrepancy (ΔAge) carries real biological significance. People who are predicted to be younger than their chronological age tend to have reduced mortality rates, as do their parents, and vice versa. Moreover, the entire genetic signature (GWAS Manhattan plot) for ΔAge correlates with the genetic signatures for individual mortality and for parental lifespan. Environmental factors known to influence health and longevity, like smoking, correlate with ΔAge, as do the Manhattan plots for these risk factors. Because this model was created to measure physiological aging rather than lifespan itself, it now provides a tool that can help us to better understand the aging process and the genes that influence it.

A prime goal of this study was to identify genes that might influence the intrinsic human rate of aging. Different species have remarkably different lifespans, but they all arose from a common evolutionary precursor. Therefore, the genetic contribution to aging is very large. Consistent with this, single-gene mutations can have dramatic effects on the overall rates of aging of model organisms. Several recent studies have shown that genetic differences only account for about 10% of the differences in longevity *within* the human population. The simplest explanation for this is that the genes that make us age so much slowly than dogs do not differ substantially from one human to another.

Unlike previous models, this UKBB ΔAge model incorporates direct physiological information to allow the identification of genes that might influence human aging independently of their effects on lifespan. Moreover, by identifying genes whose inclusion is not dependent on the presence of traits specific for any one phenotypic cluster (organ system), and that can still be detected in the UKBB healthy-only population, we can increase the chances of identifying genes that affect the intrinsic rate of aging fundamentally; that is, potentially, the genes that makes us all age more slowly than dogs. Conversely, the genes that do drop out

might be novel organ-specific disease genes or new genes influencing one or more age-related diseases.

We found that the genetic contribution to ΔAge was ~10%, similar to that for longevity, and moreover that many of the genes we identified are consistent hits in longevity GWAS analysis. Intuitively, this would be expected, since aging is a risk factor for death. However, our model allows one to dig deeper and ask whether a longevity GWAS locus might be identified only because alleles prevent advancement to extremely old age. One could imagine that this is the case for *APOE*, since *APOE4* individuals generally die prematurely of Alzheimer's disease [124]. However, we find that at all ages, even as young as 40 years, *APOE* genotype influences ΔAge, perhaps due to its more general effects on lipid homeostasis [25] or inflammation.

Pathway analysis of our GWAS hits identified neuronal function as a potential determinant of human aging, consistent with the fact that environmental factors associated with cognition, such as level of education, and a new trait, the tendency to play computer games, associate with ΔAge. How, molecularly, neuronal function might influence the aging process remains a fascinating unanswered question.

It is intriguing to consider the basis for the association of computer-game playing with youthfulness. Real-time strategy games are known to improve information processing abilities [11,12] and memory [13]. And therefore, it is possible that people who play computer games benefit from this activity because it improves brain function (note that "time to a match" is a heavily-weighted component of our biological age index; Fig. 2e, f). Computer gaming could also engender social affiliation (with other gamers), a correlate and potential determinant of longevity. Alternatively, it is possible that a genetically-defined trait, such as enhanced (or inhibited) synaptic activity in the brain or periphery could influence physiological aging, and also traits like competitiveness or adventure seeking that could predispose people to play computer games. In this case, the genetic signature that predisposes people to play games often would be the same as slow-aging signature. It is interesting to note the effect is much stronger in men (Supp. Fig. 2a) than in women (Supp. Fig. 2c), where the trend is only observed in older participants. The reason is not clear and might involve the type of games that people of different ages and genders play. It would be an interesting venue of research to document and analyze if types of games: social, like online multiplayer games, strategy-heavy, like chess, or reaction time – dependent, like first person adventures have associations with health and biological age.

The rich variety of age-dependent traits measured in the UKBB study allows a new systems-level analysis of aging at the physiological and/or molecular level, though phenotypic clustering. Some pairings, like left and right-hand grip strength are intuitively obvious, whereas others, like vitamin D and sleep duration, are not intuitive at all. *Post hoc* assessment of this latter association can be explained in the light of previous medical studies, and we predict that investigating analogous non-intuitive clusters might lead to a new understanding of causal relationships as well.

When analyzing the many phenotypes that predict aging using PLS modelling, we discovered that only 9-11 axes are necessary to predict age. This suggests that there are might be only ~10 independent systems (physiological networks) driving human aging. Interestingly, although overall the traits that figure most prominently into the sum of the principal components tend to map onto individual phenotypic clusters ("dropout clusters"), together the "meaning" of the differentially weighted sets that comprise each principal component is not obvious. For example, the 10 dropout clusters we used are not representative of the 10 axes identified in PLS analysis. It would be interesting to understand the physiological significance of each axis to better understand the process of aging. That said, the two most valuable phenotypes used in our study (those that had, overall, the most weight in age prediction) were forced vital capacity and blood pressure. Moreover, the genetic signature (GWAS Manhattan plot) of ΔAge was similar to the genetic signature for FVC and blood pressure (figure 4e). These phenotypes are integrated, multidimensional health measurements. Using genetic information to better understand age-related phenotypes through PLS axis decomposition might be a fruitful direction for future research. Finally, from a practical perspective, we suggest that measuring human biological age using the 12 simple but diverse physiological measurements that together capture ~87% of the full ΔAge model (systolic blood pressure, forced expiratory volume, and so forth; see Fig. 2), might have actuarial and clinical value. For example, this physiological-age index could be measured longitudinally to learn how aging trajectories might be affected by environmental factors or anti-aging therapeutics.

Author Contributions

S.L. and C.K. conceived the idea for the study and designed experiments, S.L. performed modelling and calculations, S.L. and C.K. wrote the manuscript, A.C. maintained and troubleshooted computing clusters and scripts necessary for data acquisition and GWAS.

Acknowledgments

Authors are grateful to the Calico community for support and discussions. Specifically, we are grateful to Eugene Melamud and his group for their support of the UKBB data framework, help with data processing and analysis, and numerous constructive discussions; we are grateful to Aarif Khakoo for his insights into disease-vs-aging paradigms; Kevin Wright and Graham Ruby for their discernment of human genetics of aging; and to David Botstein for his insights into statistical interpretation of our computational results and future applicability of such analysis. We are grateful to Madeleine Cule for supporting Calico's UKBB data interface and GWAS cluster maintenance and for guidance with GWAS methodology. We are grateful to Amoolya Singh for providing guidance in regression modelling, statistical analysis and interpretation of the data. The authors are grateful to Jonathan K. Pritchard for constructive discussions about modelling, genetic analysis, and results interpretation. This study was carried out using UK Biobank Application number 44584, and we thank the participants in the UK Biobank study. This study was funded by Calico Life Sciences LLC.

Declaration of Interests

All coauthors work for Calico Life Sciences LLC, a pharmaceutical company engaged in understanding the biology of aging and development of therapies that would ameliorate suffering from age-associated diseases.

References

1. Wright, K. M. *et al.* A prospective analysis of genetic variants associated with human lifespan. *G3: Genes, Genomes, Genetics* **9**, 2863–2878 (2019).
2. Graham Ruby, J. *et al.* Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics* **210**, 1109–1124 (2018).
3. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging* **9**, 2504–2520 (2017).
4. Min, K. J., Lee, C. K. & Park, H. N. The lifespan of Korean eunuchs. *Current Biology* vol. 22 R792–R793 (2012).
5. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology* **16**, 25 (2015).
6. Chen, W. *et al.* Three-dimensional human facial morphologies as robust aging markers. *Cell Research* **25**, 574–587 (2015).
7. Kirkwood, T. B. L. Deciphering death: a commentary on Gompertz (1825) 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies.' *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140379 (2015).
8. Milman, S. & Barzilai, N. Dissecting the mechanisms underlying unusually successful human health span and life span. *Cold Spring Harbor Perspectives in Medicine* **6**, (2016).
9. Chetty, R. *et al.* The association between income and life expectancy in the United States, 2001-2014. *JAMA - Journal of the American Medical Association* **315**, 1750–1766 (2016).
10. Wang, J. & Geng, L. Effects of socioeconomic status on physical and psychological health: Lifestyle as a mediator. *International Journal of Environmental Research and Public Health* **16**, (2019).
11. IL, W., A, P. & S, V. Intensive video gaming improves encoding speed to visual short-term memory in young male adults. *Acta psychologica* **142**, 108–118 (2013).
12. CS, G., MA, S., K, M., E, K. & None, D. B. The effect of action video game experience on task-switching. *Computers in human behavior* **28**, 984–994 (2012).
13. Waris, O. *et al.* Video gaming and working memory: a large-scale cross-sectional correlative study. *Computers in human behavior* **97**, 94 (2019).
14. Kunert-graf, J., Sakhanenko, N. & Galas, D. Allele Frequency Mismatches and Apparent Mismappings in UK Biobank SNP Data 2 Minor Allele Frequencies. 1–9 (2020).
15. Shepherd, M., Shields, B., Ellard, S., Rubio-Cabezas, O. & Hattersley, A. T. A genetic diagnosis of HNF1A diabetes alters treatment and improves glycaemic control in the majority of insulin-treated patients. *Diabetic Medicine* **26**, 437–441 (2009).
16. Abel, E. V. *et al.* HNF1A is a novel oncogene that regulates human pancreatic cancer stem cell properties. *eLife* **7**, (2018).

17. Xu, M. *et al.* Apolipoprotein E Gene Variants and Risk of Coronary Heart Disease: A Meta-Analysis. *BioMed Research International* **2016**, (2016).
18. Zheng, J. *et al.* Databases and ontologies LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
19. Melzer, D., Pilling, L. C. & Ferrucci, L. The genetics of human ageing. *Nature Reviews Genetics* vol. 21 88–101 (2020).
20. Deelen, J. *et al.* A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. *Nature Communications* **10**, 1–8 (2019).
21. van der Laan, S. W. *et al.* Cystatin C and Cardiovascular Disease: A Mendelian Randomization Study. *Journal of the American College of Cardiology* **68**, 934–945 (2016).
22. Jones, M., Denieffe, S., Griffin, C., Tinago, W. & Fitzgibbon, M. C. Evaluation of cystatin C in malignancy and comparability of estimates of GFR in oncology patients. *Practical Laboratory Medicine* **8**, 95–104 (2017).
23. Kaur, G. & Levy, E. Cystatin C in Alzheimer's disease. *Frontiers in Molecular Neuroscience* vol. 5 79 (2012).
24. Olichney, J. M. *et al.* The impact of apolipoprotein E4 on cause of death in Alzheimer's disease. *Neurology* **49**, 76–81 (1997).
25. Abondio, P. *et al.* The genetic variability of APOE in different human populations and its implications for longevity. *Genes* vol. 10 (2019).